Comment on Vul et al, ”Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition”

Thomas E. Nichols[1,2,3], Jean-Baptist Poline[2]

1 Clinical Imaging Centre, GlaxoSmithKline, London

2 Centre for Functional MRI of the Brain, University of Oxford, Oxford, UK

3 Department of Biostatistics, University of Michigan, Ann Arbor, MI

4 Neurospin, Institut d'Imagerie Biomedicale, CEA, Paris

In Press, *Perspectives on Psychological Science*.

Thomas E. Nichols, Ph.D.
Director, Modelling & Genetics
GlaxoSmithKline Clinical Imaging Centre
Imperial College
Hammersmith Hospital
London
W12 0NN
United Kingdom
thomas.e.nichols@gsk.com

Jean-Baptiste Poline, Ph.D.
Neurospin projects coordinator
Institut d'Imagerie Biomedicale, CEA
Bâtiment 145- Point Courrier 156
91191 Gif-sur-Yvette cedex
Paris
France
jbpoline@cea.fr

Abstract

Vul et al., “Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition” make a broad case that current practice in neuroimaging methodology is deficient. They go so far as to demand that authors retract or restate results, which we find wrongly casts suspicion on the confirmatory inference methods that form the foundation of neuroimaging statistics. We contend the authors’ argument is overstated and that their work can be distilled down to two points already familiar to the neuroimaging community: That the multiple testing problem must be accounted for, and that reporting of methods and results should be improved. We also illuminate their concerns with standard statistical concepts, like the distinction between estimation and inference, and between confirmatory and post hoc inferences, which makes their findings less puzzling.

We are happy that the authors have generated such a stimulating discussion over fundamental statistical issues in neuroimaging. However, the issues raised are well-known to brain imaging researchers and the paper could be distilled to two points that have already received much attention in the literature. The first one is that brain imaging has a massive Multiple Testing Problem which must be accounted for in order to have trustworthy inferences, and the presence of this problem requires careful distinction between corrected and uncorrected inferences. Second, papers in neuroimaging have methods descriptions which are confusing or incomplete, which is a disservice to scientific discourse especially as neuroimaging reaches into new applied areas.

Finding solutions to the Multiple Testing Problem (MTP) was an active area of research during the past two decades. We now have consensus methods that are widely accepted and used (see, e.g. Chapters 18-21 of Friston, 2006, or Chapter 14 of Jezzard et al, 2001). The two types of commonly used inference methods are those that control the familywise error rate (FWE; Nichols & Hayasaka, 2003) versus those that control the false discovery rate (FDR; Genovese et al., 2002). Bonferroni and Random Field Theory thresholds are two methods that control FWE, the chance of one or more false positives. A statistic image that is thresholded with a valid 5% FWE threshold is guaranteed to have *no* false positives at all with 95% confidence. FDR is a more lenient measure of false positives, and a valid 5% FDR threshold will allow as many as 5% of the suprathreshold voxels to be false positives on average. Both FWE and FDR methods can be applied voxel-

wise, as a threshold on a statistic image, or cluster-wise, as a threshold on the size of clusters after applying an arbitrary cluster-forming threshold.

Whether FWE or FDR, voxel-wise or cluster-wise, corrected inferences must be used to ensure that results are not attributable to chance. Such corrected inferences are known as “confirmatory” inference, a test of a pre-specified null hypothesis with calibrated false positive risk. This is in distinction to exploratory or “post hoc” inference, where no attempt is made to control false positive risk. Reporting and interpreting the voxels or clusters that survive a corrected threshold is a valid confirmatory inference and the foundation of brain imaging methodology. Complete reporting of these results usually consists of a corrected P and raw t (equivalently r) value, and in no way does the unveiling of the t value invalidate this inference.

The authors suggest that that the raw t (equivalently r) scores that survive a corrected threshold are "impossible"; this is incorrect because they are simply local extremal values that should be reported for what they are, post hoc measures of significance uncorrected for multiple testing. Crucially, as the raw scores are uncorrected measures, they are *incomparable* with a behavioral correlation that did not arise out of a search over 100,000 tests.

This incompatibility issue is also related to how the authors misinterpret the reliability result (Nunnally, 1970), applying it to *sample correlations* when it is statement about *population correlations*. There is in fact substantial variation in a sample correlation about its true population value, with the approximate

standard error of r being $1/\sqrt{n}$. Thus a sample correlation based on 25 subjects has an approximate 95% confidence interval of ± 0.4, and indicates that, in this setting, an r of 0.9 is entirely consistent with a ρ of 0.7 and is indeed "possible".[1]

The second essential point of the paper is that publications in neuroimaging have methods descriptions which are confusing or incomplete. While this is a point of embarrassment for the field, it is a point that has been addressed in several publications (Poldrack et al, 2007; Carter et al, 2008; Ridgway et al, 2008). If there is any misdeed committed by the "red" papers, perhaps it is that they failed to fully label the inferences as post hoc. That extremal-selected post hoc tests give rise to greater correlations than confirmatory tests (Figure 5 in Vul et al) is self-evident and not worthy of the tenor of the note. In particular, we argue that, while authors have the responsibility to clearly and completely describe their methods and results, readers have the responsibility to understand the technology used and how to correctly interpret the results it generates. For example, in the field of genetics, whole-genome association analyses search over 100's of 1,000's of tests for genotype-phenotype correlations and publications routinely include plots of uncorrected P-values (see, e.g., Fig. 4 in (The Wellcome Case Control Consortium, 2007)). Yet we are unaware of any movement to suppress these plots from publication, presumably because genetic researchers understand the difference between these massive analyses and candidate Single Nucleotide Polymorphism analyses where no multiplicity is involved.

[1] More accurate confidence intervals computed with Monte Carlo or Fisher's Z transformation will be shorter than ±0.4, but still make the point of substantial sampling variability about the population value.

We must also take issue with the seemingly most compelling argument of the paper, based on Figure 4 and the weather-stock market correlation example. The problem with these examples is that they are based entirely on a null-hypothesis argument, i.e., the total noise case. However, if the papers examined use corrected thresholds, then the suprathreshold voxels will be mostly or entirely true positives.

As reviewed above, a 0.05 voxel-wise FWE threshold guarantees no more than a 0.05 chance that any null voxels will survive the threshold. In this case, Figure 4(a) is totally irrelevant (with 95% confidence) and the distribution of suprathreshold correlations is purely due to true positives. If, instead, the papers cited use a 0.05 FDR threshold, the suprathreshold voxels will be a mixture of true and false positives, but the fraction of false positive voxels will be no more than 5% on average.

Finally, we find that the focus on correlation itself is problematic, as the correlation coefficient entangles estimation of effect magnitude and inference on a non-zero effect. A much more informative approach is to separately report significance and effect magnitude. That is, report significance with a corrected P-value and report effect magnitude (still post hoc, of course) with a unit change in social behavioral score per unit percent BOLD change (as recommended in Poldrack et al., 2007). The behavioral scores have known scales and properties, and the percent BOLD change has an approximate interpretation of percent change in blood flow (Moonen et al, 2000).

Reporting such measures will provide more interpretable and comparable measures for the reader.

The authors seem to be arguing that the field of neuroimaging should turn away from *inference* on where an effect is localized, and focus instead solely on *estimation* of effect magnitude assuming a known location (Saxe et al., 2006). This is a significant shift in perspective that justifies ample and perhaps strident scientific discourse but not bluster that suggests standard inferential practice is fraudulent.

We would like to thank the authors for an engaging article that raises issues that apply to every neuroimaging study. However we maintain that the community would have been better served if the alarmist rhetoric had been replaced by a measured discussion which made connections to standard statistical practice, distinguishing between estimation and inference, and between confirmatory and post hoc inferences, and had simply acknowledged the incomparability of reported post hoc imaging correlations with other correlations in the psychology literature.

Acknowledgement

The authors would like to thank Michael Lee for input on this comment.

References

Carter, CS, Heckers, S, Nichols, T, Pine, DS, & Strother, S (2008). Optimizing the Design and Analysis of Clinical FMRI Research Studies. *Biological Psychiatry*, 64(10), 842-849.

Friston, KJ (2006). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Amsterdam, The Netherlands: Academic Press.

Jezzard, P, Matthews, PM, Smith, SM (Eds.) (2001). *Functional MRI: An Introduction to Methods*. Oxford, United Kingdom: Oxford University Press.

Genovese, CR, Lazar, N, Nichols, TE. (2002). Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage*, 15, 870-878.

Moonen, CTW, Bandettini, PA (2000). *Functional MRI*. Berlin, Germany: Springer.

Nichols, TE and Hayasaka, S. (2003). Controlling the Familywise Error Rate in Functional Neuroimaging: A Comparative Review. *Statistical Methods in Medical Research*, 12(5), 419-446.

Poldrack, PC, Fletcher, RN, Henson, Worsley, KJ, Brett, M, Nichols, TE. (2007). Guidelines for reporting an fMRI study. *NeuroImage*, 40(2), 409-414.

Ridgway, G. R.; Henley, S. M. D.; Rohrer, J. D.; Scahill, R. I.; Warren, J. D. & Fox, N. C. (2008). Ten simple rules for reporting voxel-based morphometry studies. *NeuroImage*, 40, 1429-1435.

Saxe, R, Brett, M, Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. NeuroImage 30(4), 1088-1096.

Vul, E, Harris, C, Winkielman, P, Pashler, H. "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition". *Perspectives on Psychological Science.* (In Press).

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447, 661-678.